智能 N 维向量的空间模型

史德琴[1]，余　莉[2]
(1. 空军工程大学 工程学院，陕西 西安 710038；2. 94669部队 自动化站，安徽 芜湖 241000)

摘　要：传统向量空间模型在计算复杂度、查询性能、智能性方面存在种种缺陷。在其基础上，提出了智能 N 维向量空间模型，改进了文档特征向量生成的算法，使用局部统计数据计算特征向量，大大降低了计算复杂度。模型采用用户点击作为反馈，提出了对初始的文档特征向量和用户查询向量进行调整的算法。最后，从理论上对两种模型的性能进行了对比分析。

关键词：信息获取；信息检索；向量空间模型；智能向量空间模型；WEB检索
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-3516(2004)03-0083-04

当前，搜索引擎(Search Engine)是在WEB上寻找信息的最主要工具，一般利用机器人(Robot)以一定的策略收集WEB文档，对信息进行理解、提取、组织和处理，为用户提供信息检索服务，从而起到网络导航的目的。这种对信息进一步处理的方法就是信息获取模型(Information Retrieval Model)[1]研究的主要内容。本文在对现有的主要模型进行对比分析的基础上，提出了采用局部统计数据生成文档特征向量，并利用用户反馈对文档特征向量和查询向量进行调整的智能 N 维空间向量空间模型。

1　信息获取模型

信息获取模型主要研究文档和查询的表示方式以及查询与文档的匹配方法。目前，主要的信息获取模型有布尔模型、向量空间模型、概率模型、神经网络模型、聚类模型、基于规则的模型、模糊模型和语义模型等。其中，又以布尔模型和向量模型的研究最为成熟。

1.1　布尔模型

在构造信息检索系统时，布尔模型使用的最普遍的，特别是在商用信息检索系统中。在布尔模型中，用一个来自词典的关键词集合来表示一个文档，用通过逻辑操作符(与、或、非)联结的关键词来描述查询，在对查询进行文档匹配时，主要通过集合运算，来判断该文档是否满足查询的条件。

1.2　向量空间模型

向量空间模型则是在实验环境中研究最多的模型[2-3]。模型将给定的文档和查询转换成一个特征向量，其最大特点是可以方便地计算出两个向量的近似度，既向量所对应的文本相似性。所有文档用向量表示，也就是将搜索到的文档材料进行关键词抽取，形成特征向量，而用户查询时，则针对特定的查询向量，比较它与所有文献的相似度，并依相似度大小将文献排序提交给用户。

1.2.1　文档特征向量

在文档集合 D 中的文档 d_i 可以表示为特征向量($x_{i1}, x_{i2}, \cdots x_{in}$)，其中的权值 x_{ij} 描述了关键词 t_j 代表文档 d_i 能力的大小。如果关键词 t_j 和文档 d_i 的相关性越强，就越能代表该文档，权值 x_{ij} 也应越大：

$$x_{ij}=\frac{t_{f_{ij}}}{d_{f_j}}=t_{f_{ij}}id_{f_j}=t_{f_{ij}}\left(\log_2\frac{K}{n_j}+1\right) \tag{1}$$

式中：K 代表文档集合 D 中的文档个数，n_j 代表在文档集合中出现关键词 t_j 的文档数目。

收稿日期：2004-01-15
作者简介：史德琴(1970-)，女，河北涿鹿人，讲师，硕士生，主要从事网络安全研究.

$t_{f_{ij}}$(Term Frequency)指某个关键词 t_j 在文档 d_i 中出现的频率。$t_{f_{ij}}$越大,表示文档 d_i 包含关键词 t_j 的次数较多。从式(1)可知,$t_{f_{ij}}$越大,x_{ij}值越大;同样 n_j 越小,x_{ij}值也越大,说明该关键词 t_j 更能够代表文档 d_i 的内容。d_{f_j}(Document Frequency)指整个文档集合 D 中,包含关键词 t_j 的文档个数。直观上看,d_{f_j}越大,即包含关键词 t_i 的文档越多,则关键词 t_j 代表文档 d_i 的能力就越小,t_j 的权值应当与之成反比。为了计算的方便,通常使用 IDF(Inverse Document Frequency),IDF 与 t 的权值成正比。TF 的计算只需要单个文件中关键词的统计信息,而 DF 和 IDF 的计算则需要对整个文档集合进行统计,计算量较大。

1.2.2 查询特征向量

查询特征向量的生成一般采用布尔框架。在布尔框架中,t 的权值要么是 1,要么是 0。如果在查询字符串 QS(Query String)中包含关键词 t,则 t 的权值为 1,否则为 0。查询条件的向量化见式(2)。除了布尔框架以外,查询向量中关键词的权值也可以由用户指定。

$$y_j = \begin{cases} 1 & \text{如果 } t_j \in \text{QS} \\ 0 \end{cases} \tag{2}$$

2 智能 N 维向量模型

本文在传统 N 维向量空间模型的基础上,提出了智能 N 维向量空间模型。智能 N 维向量空间模型采用了传统 N 维向量空间模型的基本框架,继承了传统 N 维向量空间模型的相似函数,着重对文档特征向量和查询特征向量的生成算法进行了改进,利用用户点击作为反馈,通过机器学习对文档特征向量和查询特征向量进行调整,使模型具有一定的智能性。

模型定义:设文档集合 D 中共有 N 个不同的关键词 t,则 D 中的任意文档 d_i 可以用一个特征向量来表示,$d_i = (x_{i1}, x_{i2}, \cdots x_{in})$,其中 x_{ij}表示第 j 个关键词 t_j 对文档 d_i 的权值;查询字符串 $Q = (y_1, y_2, \cdots, y_n)$,其中 y_j 表示第 j 个关键词 t_j 对查询的权值;文档 d_i 和查询 Q 的相似性表示为特征向量的相似函数 $\text{Sim}(d_i, Q)$。

$$\text{Sim}(d_i, Q) = \cos(d_i, Q) = \left(\sum_{j=1}^{N} x_{ij} y_j\right) \Big/ \sqrt{\sum_{j=1}^{N} x_{ij}^2 \sum_{j=1}^{N} y_j^2} \tag{3}$$

2.1 基于关键词属性的向量生成

传统向量空间模型认为文档中所有关键词的描述能力是相同的,这种过于简单的假设缺点很明显。例如,在一个文档集中有四个文档 d_1、d_2、d_3 和 d_4,这四个文档中都包含关键词 t,并且 t 在这四个文档中出现次数都为 k 次。但是,在文档 d_1 中,t 出现在标题中,在文档 d_2 中,t 是出现在摘要中,在文档 d_3 和 d_4 中,t 是出现在正文中。显然,出现在标题中的关键词要比出现在摘要中的更能确切代表文档的内容,同样出现在摘要中的关键词也要比出现在正文中的更能代表文档的内容;而对于文档 d_3 和 d_4 来说,如果 d_3 的正文长度大于 d_4 的正文长度,则相比之下也可以认为 t 对于文档 d_4 而言它所能代表文档内容的能力要比 d_3 强。如果采用关键词 t 进行查询,理想中的检索模型应该检索结果按照 d_1、d_2、d_4 和 d_3 降序排列,而如果使用传统 N 维向量空间模型,对各个文档特征向量的计算结果是完全相同的,因而给出的检索结果也是无法预知的。

如果在特征向量的构造过程中,能够利用有助于判断关键词描述能力的其它信息,就能够避免这种情况的发生。设在文档集合 D 中共有 K 篇文档,N 个关键词,则文档 d_i 的特征向量为$\{x_{ij}\}$,算法定义如下:①划分文本段。根据统一的标准,将文档集合 D 中的所有文档作 M 种划分。例如:标题、小标题、摘要、正文等。②计算文本段权重。根据重要程度,计算每一种文本段 S_k($k \leqslant M$)的权重 W_{sk}($k \leqslant M$)。例如:标题 =4,小标题 =3,摘要 =2,正文 =1;③利用公式(4)计算关键词权值。

$$x_{ij} = \sum_{k=1}^{m} N_{j,sk} W_{sk} / l_{sk} l_i \tag{4}$$

式中:$N_{j,sk}$表示关键词 t_j 在文本段 S_k 中出现的次数;l_{sk}代表该文本段的长度,l_i 代表文档的总长度。

采用上述算法对上一小节中提到的文档的特征向量计算,见表 1,设 d_1、d_2、d_4 的文档总长度为 10 + 100 + 1 000 = 1 110,d_3 的文档总长度为 10 + 100 + 1 500 = 1 610。

显然,$X_{1j} > X_{2j} > X_{4j} > X_{3j}$,该排序结果是符合一般预期的。该算法中采用了关键词在文档中的位置属性。其实,在对文档进行关键词提取的过程中,在统计关键词的出现次数的同时,提取其他的相关信息也是十分方便的(只需一次文档扫描),如:关键词位置、字体大小、字形粗细、字体名称、是否下划线、是否斜体等

等。使用这些属性的算法基本一致,也可以同时采用多种属性,只要为不同的属性定义不同的权值即可。

表1　位于文档不同位置的关键词权值计算表

文档	出现次数	出现位置	W_{ik}	l_{ik}	l_i	X_{ij}
d_1	K	标题	4	10	1 110	$X_{1j}=4K/11\ 100$
d_2	K	摘要	2	100	1 110	$X_{2j}=2K/111\ 000$
d_3	K	正文	1	1 500	1 610	$X_{3j}=K/2\ 415\ 000$
d_4	K	正文	1	1 000	1 110	$X_{4j}=K/1\ 110\ 000$

在算法的实现上,由于x_{ij}的计算结果一般是浮点数,为了节约存储空间,可以将向量进行分级表示,如将X_{1j}、X_{1j}、X_{3j}、X_{4j}分别设为第3、2、1、0级,这样只需要2个比特就可以表示一个关键词的向量。K个文档的N维特征向量的存储在理论上需要$NK/4$个字节。级别分得越细,描述向量之间差异的能力就越强,同时消耗更多的存储空间,以及在相似性比较时更多的计算时间。

2.2　基于反馈的文档特征向量调整

反馈是控制论中的重要手段,用输出来调整系统,调节系统中不稳定的因素。在信息检索中,反馈一样可以发挥这样的作用。在特征向量的生成中,无论采用什么方法,都是对文档内容的一种猜测。因而不精确性是无法避免的,基于用户的反馈则可以减少这种不精确性。

用户在利用关键词进行检索时,检索系统返回相关的文档集合。如果用户点击了返回集合中的某篇文档,则认为用户对该关键词表示该篇文档的能力表示认可,一次点击代表一次投票V,则文档的特征向量就可以根据用户的投票数进行调整。

$$x_{ij,\text{new}} = x_{ij,\text{old}} + \alpha\sum_t V_{ij} - \beta\sum_t \overline{V}_{ij} \tag{5}$$

式中:$x_{ij,\text{old}}$表示关键词t_j对文档d_i的特征向量初始权重;$x_{ij,\text{new}}$表示调整后的权重;$\sum_t V_{ij}$表示在一段时期t内,用户对t_j和文档d_i的肯定投票数,$\sum_t \overline{V}_{ij}$代表用户的否定票数;t代表特征向量调整的频繁程度,可以考虑以"天"为单位;α、β是调整系数,反映了向量调整的灵敏度,系统使用初期可以考虑用$(\sum_t V)^{-1}$,即每段时间内总点击数的倒数。随着系统使用时间的增加,需要根据实际使用情况进行调整,调整原则是保持$\sum_{i=1}^{K}\sum_{j=1}^{N} x_{ij}$基本为常数。

向量调整的算法如下:① 根据查询向量返回检索结果。② 收集用户投票数。假设用户用关键词tj查询,检索系统返回K个文档,如果用户点击了其中的K_1个文档,则在该K_1个文档中,对关键词t_j的肯定投票数加1,其它$(K-K_1)$个文档中对关键词tj的否定投票数加1。③ 收集每天的总点击数,计算α、β。④ 根据公式(5)调整文档的特征向量。

2.3　利用关键词的相关性调整查询向量

为了能够在一次查询中就能够得到较满意的结果,本文提出利用过往的查询来对当前的查询向量进行调整。假设用户使用关键词向量$\{t_i\}$来构造一个查询,在系统返回的结果中,如果用户选中了某篇文档,则认为所有包含在该文档中的关键词t两两之间具有强关联. 一次点击代表该用户对该文档中包含的所有关键词两两之间的强关联都投了一次肯定投票V。

$$y_{ij,\text{new}} = y_{ij,\text{old}} + \alpha(\sum_t V_{ij})/N_{\text{click}} \tag{6}$$

式中:$y_{j,\text{old}}$表示调整前的查询向量;$y_{j,\text{new}}$表示调整后的查询向量。$y_{j,\text{old}}$可以采用其他经典的计算方法获得,如公式(2)。$\sum_t V_{ij}$表示在一段时期t内,用户对向量R_i和R_j之间强关联关系的肯定投票数。N_{click}表示该段时期内用户的总点击数。t代表特征向量调整的频繁程度,可以考虑以"天"为单位。

α是灵敏度系数,α取值太高容易使包含弱相关关键词的文档排序过于靠后,该值应该随着系统的使用动态调整。初期可以考虑如下估计公式(n为查询关键词的个数):

$$\alpha = (\max(y_j) - \min(y_j))/n \tag{7}$$

查询向量生成的基本过程如下:① 利用经典算法计算$y_{j,\text{old}}$;运用公式(7)计算$y_{j,\text{new}}$;② 计算与文档特征向量的相似性,返回检索结果;③ 收集用户的点击,计算N_{click}和$\sum_t V_{ij}$;④ 等待下一次检索。

在一个包含K个关键词的系统进行包含n个关键词的查询,查询向量调整的时间复杂度[4]是$O(n^2)$,由于n一般较小,所以采用上述算法带来的时间延迟是可以接受的。

上述调整算法实质上也是一种反馈机制,只不过该反馈机制利用了其它用户的知识对相关性的判断知识,可以在一次查询中就获得比较好的查询向量,代价是最多需要 K^2 个空间单位来存储关键词之间的关系。这种基于专家知识的反馈机制使得查询向量的构造也具有的一定的智能性。

2.4 性能分析

传统 N 维向量空间的权值计算公式中的 n_j 是一个全局的统计数据,需要对所有收集到的文档进行统计以后才能得出。在一个文档个数为 K 的文档集合中,计算一个文档的 N 维特征向量,其时间复杂度为 $O(NK)$,对所有 K 个文档计算特征向量,时间复杂度为 $O(NKK)$。智能 N 维向量空间模型计算一个文档的 N 维特征向量,时间复杂度为 $O(N)$,计算包含 K 个文档的文档集合的特征向量库,总的计算时间复杂度为 $O(NKK)$。

传统 N 维向量空间模型对出现在文档任何位置的所有关键词一视同仁,显然有悖于人们的文档撰写习惯。智能 N 维向量空间模型考虑到不同位置的关键词具有不同的描述能力,为每种文本段赋予不同的权值,并认为关键词的描述能力与所在文本段的长度成反比,因而对半结构化文档的特征具有更好的描述性能,且符合文档构成的一般规律。

传统模型无法利用用户点击的反馈信息,特征向量一旦生成,就不再改变。特征向量的完全取决于初始的生成算法,描述描述质量难以得到提高。智能 N 维向量空间模型则充分利用可以获得的用户的反馈,通过用户的投票对特征向量进行调整,使得特征向量的描述质量得到不断的改进提高。

传统模型的查询向量生成基于固定的算法,造成用户构造查询的困难。智能 N 维向量空间模型则利用用户的点击作为反馈,通过不断学习来判断关键词之间的相关性,并对初始的查询向量进行权值调整。使得一般用户也可以借助于其它用户的知识,得到更为满意的查询结果。

3 结论

智能 N 维向量空间模型在传统向量模型的基础上,改进了文档特征向量生成的算法,使用局部统计数据计算特征向量,大大降低了计算复杂度,对于降低检索系统索引库的建设成本具有突出的意义。同时,模型采用用户点击作为反馈,对初始的文档特征向量和用户查询向量进行调整,在理论上可以改进向量描述的质量,提高信息检索的满意程度。

参考文献:

[1] LANCASTER F W. Information Retrieval Systems: Characteristics, Testing and Evaluation[M]. New York: Wiley, 1968.

[2] Salton G, Wang A, Yang C. A Vector Space Model for Information Retrieval[J]. In Journal of the American Society for Information Science, 1975, 18: 613 - 620.

[3] Brin S, Page L. The anatomy of a large - scale hypertextual Web search engine[A]. In Proceedings of the 7th International World Wide Web Conference[C]. Brisbane, Australia, 1998.

[4] 王晓东. 计算机算法设计与分析[M]. 北京:机械工业出版社,2001.

(编辑:姚树峰)

Intelligent N - Dimension Vector Space Model

SHI De - qin[1], YU Li[2]

(1. The Engineering Institute, Air Force Engineering University, Xi'an, Shaanxi 710038, China; 2. Automation Station of Unit 94669, Wuhu, Anhui 241000, China)

Abstract: There are disadvantages of traditional vector space model in computational complexity, query efficiency and intelligence. Based on the traditional vector space model, an intelligent N - dimension vector space model (INDVSM) is proposed. INDVSM shows distinct improvement in the computational complexity of document vectors by using a refined algorithm based on local information, and it takes the user's clicks as feedback to tune the initial vectors of documents and queries. Meanwhile the tuning algorithm is also presented and these two models are compared theoretically.

Key words: information retrieval; vector space model; intelligent N - Dimension vector space model; WEB information retrieval